BOOK II

Collaborating Project Reports and Supporting Appendixes

The following sections detail the reports and plans for SUMEX-AIM collaborating projects and also contain additional information in the form of appendixes relating to the core resource progress and operation. The heading and page numbering of these sections does not continue sequentially from that of the Book I progress report. The discontinuity reflects the initial organization of this material as part of our renewal grant application.

6 COLLABORATIVE PROJECT PROGRESS AND OBJECTIVES

The following subsections report on the collaborative use of the SUMEX facility including the formally authorized projects within the Stanford and AIM aliquots and the various "pilot" efforts currently under way. These project descriptions and comments are the result of a solicitation for contributions sent to each of the project Principal Investigators requesting the following information:

I) Summary of research program
   A) Technical goals
   B) Medical relevance and collaboration
   C) Progress summary
   D) Up-to-date list of publications

II) Interactions with the SUMEX-AIM resource
   A) Examples of collaborations and medical use of programs via SUMEX
   B) Examples of sharing, contacts and cross-fertilization with other SUMEX-AIM projects (via workshops, system facilities, personal contact, etc.)

We believe that the reports of the individual projects speak for themselves as rationales for participation; in any case the reports are recorded as submitted and are the responsibility of the indicated project leaders.

6.1 STANFORD PROJECTS

The following group of projects is formally approved for access to the Stanford aliquot of the SUMEX-AIM resource. Their access is based on review by the Stanford Advisory Group and approval by Professor Lederberg as Principal Investigator. As noted previously, the DENDRAL project was the historical core application of SUMEX. Although this is described as a "Stanford project," a significant part of the development effort and of the computer usage is dedicated to national collaborator-users of the DENDRAL programs.

6.1.1 DENDRAL PROJECT

DENDRAL - Resource Related Research - Computers & Chemistry

Carl Djerassi, Principal Investigator
Professor of Chemistry
Stanford University

I. OVERVIEW OF RESEARCH ACTIVITIES

Technical Goals

Our research, development and future plans focus on both the question of structure elucidation in general and the problem of providing computer assistance to scientists engaged in specific aspects of this important activity.

A simplified representation of major milestones in solving unknown biomolecular structures by manual methods is presented in Figure 1.

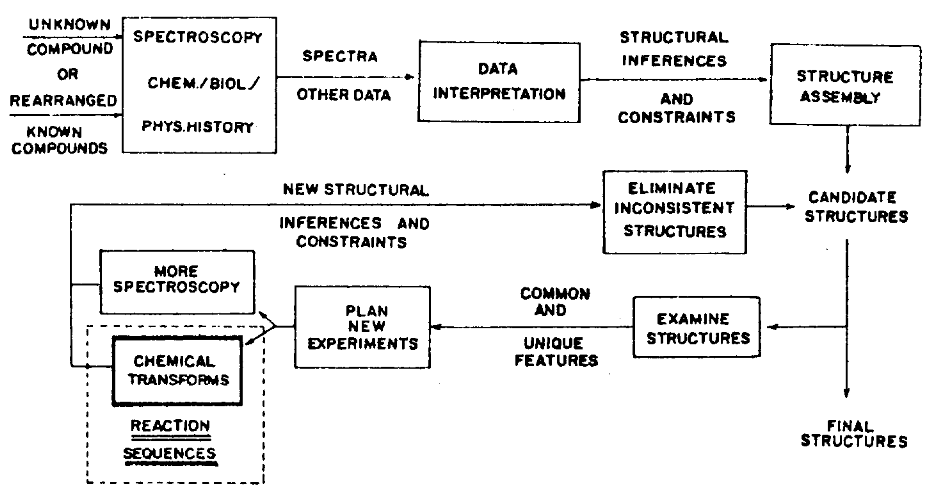


Figure 1. Important steps in manual solution of structures of unknown chemical compounds.

These steps, indicated as separate boxes, may be performed explicitly or implicitly. There are considerably more complex relationships among the boxes of Fig. 1 than are indicated when structures are actually solved. Nevertheless, the Figure provides a good introduction to both our recent work and our future directions. We describe briefly each of the milestones in the following paragraphs. More detailed discussions of each topic follow in subsequent sections.

The first step in identification of an unknown structure is to separate it from other components in a potentially complex mixture and to isolate it in reasonably pure form. These steps are performed by scientists, frequently with the assistance of various instruments. Although our research is not directed toward any part of this separation and isolation procedure (except insofar as these procedures also yield data which are subject to computer-assisted interpretation), information about the chemical and physical characteristics of the compound may be crucial to further efforts to determine its structure.

Depending on the quantity of sample available and its characteristics, various spectroscopic and additional chemical data are then collected on the unknown. A mass spectrum is frequently obtained, e.g., from a combined gas chromatograph/mass spectrometer (GC/MS) system. An important part of our recent proposal to the NIH is directed toward automation of combined GC/MS systems operated at high mass spectrometer resolving powers. Data on elemental compositions and relative ion abundances are then available in computer-readable form for further analysis (see MSRANK). The chemist possess an armamentarium of spectroscopic techniques which can be brought to bear on a structure. One advantage of our work is that any data so obtained can be used to help solve the structure as long as it can be expressed, manually or by computer, in substructural statements about the unknown.

The next important phase in structure elucidation is interpretation of the available data (Fig. 1) in terms of structural features of the molecule. These interpretations may be in terms of known structural units ("superatoms", polyatomic aggregates of atoms in known configurations), or in terms of structural units, ring sizes, proton or carbon distributions. The latter set of features represents constraints on the kinds of structures which are possible. Our efforts in the area of computer-assisted data interpretation are focussed on mass spectral and carbon-13 nuclear magnetic resonance (13CMR) data. We are developing general approaches to automated analysis of these data in terms of structural features of unknowns.

Our recent efforts are summarized in Figure 2, and discussed in detail subsequently. We have been concerned with use of these data from two points of view, planning and prediction (Fig. 2). During planning, experimental data are examined in order to extract specific structural information to be used in assembling candidate structures. In prediction each candidate structure is tested to determine how closely its predicted spectrum agrees with the observed spectrum. The candidates can be ranked accordingly. The Meta-DENDRAL research is directed toward determination of rules of spectroscopic data which can be used either for planning or prediction (see below).

Given possible structural fragments of the complete molecule and constraints on how these fragments may be assembled into complete molecules, a process of structural assembly follows (Fig. 1). There has been no proven algorithm for solving this problem prior to earlier work supported by the current grant. Traditionally, this process has been left to manual, pencil and paper work. Our CONGEN program, which was designed to solve this problem, is the farthest advanced of programs designed to assist in various aspects of structure elucidation. It performs the structural assembly process, under constraints, and

DATA INTERPRETATION

"PLANNING"

EXTRACTION OF STRUCTURAL INFORMATION DIRECTLY FROM SPECTROSCOPIC DATA.

1. MASS SPECTRA - MDGGEN
2. 13CNMR

PREDICTION

USE OF SPECTROSCOPIC DATA TO RANK CANDIDATE STRUCTURES.

1. MSPRUNE, MSPRED
2. 13CNMR

META - DENDRAL

FORMATION OF RULES TO BE USED FOR BOTH PLANNING AND PREDICTION.

Figure 2. Relationship between use of rules in either planning or prediction. Both approaches are used in utilizing data for structure elucidation.

allows the scientist using the program to examine structural candidates and remove those deemed implausible (Fig. 1). A large portion of our recent and future work is directed toward improving the CONGEN program and building other facilities around it (see later sections). We have demonstrated the utility of CONGEN in structural studies, and subsequent sections discuss our recent developments and applications of CONGEN as well as our interactions with other scientists desiring access to our programs.

Given a set of structural candidates, the experimenter examines them to determine what experiments might be performed to focus on the correct structure by stepwise rejection of alternative hypotheses. When there are only a small number of possibilities under consideration, manual methods suffice. But CONGEN provides the capability for exhaustive enumeration of structural possibilities at a point in a structural problem when there may be many hundreds of possibilities. It is very difficult to examine these structures and plan experiments by hand. We have begun exploring ways to provide computer assistance to this important aspect of structure elucidation. We refer to this research area as the Experiment Planner, discussed in more detail below.

When new experiments have been planned the researcher carries them out and uses the results as additional constraints on the structural candidates (Fig. 1). New experiments may include collecting of additional spectroscopic data or performing a sequence of chemical reactions on the unknown. The latter experiments may be chosen to convert the unknown into a related compound which possesses physical or chemical properties more amenable to analysis. During the past year we have developed a program to assist scientists in carrying out representations of chemical reactions in the computer and eliminating undesired structural candidates based on constraints exercised on the products of the reaction. This work is described in two subsequent sections. One section describes use of the program, which we call REACT, to explore structural possibilities exactly as outlined above. A later section describes recent progress in increasing the power of REACT.

Medical Relevance

Structure elucidation is a fundamental problem for medical practice and biomedical research. For example, we are collaborating with physicians in the Department of Pediatrics who monitor the body fluids of newborn infants in order to detect abnormal compounds. Much of the research leading to new drugs and new methods for synthesizing drugs also depends on careful analysis and identification of molecular structures of compounds. The computer tools that we are developing will aid in the determination of molecular structures by giving working scientists help with data collection, data interpretation, hypothesis testing and, most important, systematic consideration of all molecular structures that are consistent with the interpretations of the available data.

PROGRESS SUMMARY

Experiment Planner

We have begun preliminary considerations of design and implementation of an experiment planner. This program will assist chemists in designing the most effective set of experiments to perform to solve the structure. Although the experiment planner will be a future activity of our group, we are developing and using other structure manipulation functions which will provide groundwork for future developments.

One important aspect of experiment planning is the ability to examine in some way the set of candidate structures. Although many can be drawn for visual review, drawing is impractical when dozens or hundreds of structures are involved. To assist persons using CONGEN in reviewing their structures we have developed a function auxiliary to CONGEN which we call SURVEY.

SURVEY

FUNCTION: AIDS IN PERCEPTION OF ANY OF A PRE-SPECIFIED SET OF STRUCTURAL FEATURES IN A GROUP OF STRUCTURAL CANDIDATES.

E.G. A) FUNCTIONAL GROUPS
B) TERPENOID SKELETONS
C) AMINO ACID SKELETONS

Figure 3. Function of the SURVEY program and examples of recent application areas.

The function of SURVEY is summarized in Figure 3. SURVEY simply acts as a reminder to the scientist of the presence or absence of certain structures or structural features. During the past year we have used SURVEY extensively. For example, we have used it to detect implausible functional groups in a set of candidate structures, using a file of substructures representing a wide variety of functionalities. In many problems, implausible functional groups are forgotten and CONGEN is never constrained to remove them. Another example of use of SURVEY is in conjunction with collaborative work with persons in the

Department of Genetics. In analysis of serum or urinary metabolites in patients of high risk of metabolic disorder, we have had occasion to use CONGEN in exploration of unknown structures [Report HPP-77-11]. Some of these structures could formally be conjugates of amino acids with organic acids. If so, such structures will possess backbones of naturally-occurring amino acids. SURVEY was used to provide a summary of which structural candidates possessed such amino acid skeletons.

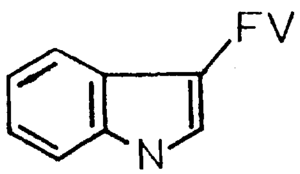
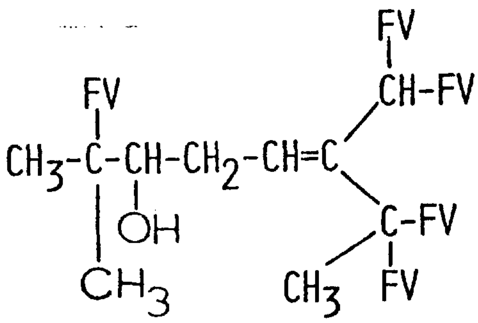
We have recently used SURVEY in a related application involving the structure of "polyalthenol", discussed by LeBoeuf, et al. (Figure 4). Superatoms and constraints supplied to CONGEN to derive structural candidates are summarized in Fig. 4.

We summarize in Figure 5 the structural possibilities which resulted. There are five structures possessing a bicyclo[2.1.1] system, and six which possess a bicyclo[4.3.1] system (Fig. 5, top). These structures are energeticaly less favorable. For example, several possess a double bond at a bridgehead atom, which violates Bredt's Rule. There remain, however, 11 structures which are not formally excluded by data presented by LeBoeuf, et al. Because these workers based their structural assignment on biogenetic grounds, we used SURVEY and REACT to test their hypothesis. We have, in computer-accessible libraries, known terpenoid ring systems which can be used within SURVEY to test sets of structures for known skeletons. None of the 22 structural candidates possesses a previously known skeleton. Because the authors postulated a relationship to a known skeleton via a single methyl shift, we used REACT to exercise a single methyl shift in all possible ways on each of the 22 candidates. SURVEY was then used to test the results for the presence of known terpenoid systems, and the drimane skeleton, the postulated precursor of polyathenol, was the only known skeleton which resulted. This does not prove the hypothesis of LeBoeuf, et al., but certainly helps strengthen it.

SURVEY is, however, only the barest beginning of an experiment planner, even though it has proven useful. We plan to build from this beginning toward a much more powerful system.

M. LeBoeuf, M. Hamonnière, A. Cavé, H. Gottleib, N. Kunesch, and E. Wenkert, Tet. Lett., 3559 (1976).

"POLYALTHENOL" $C_{23}H_{31}NO$

SUPERATOMS	ARBITRARY NAME	NUMBER
	IN	1
	BI	1
CH_3-FV	ME	1
FV-CH_2-FV	CH2	3
FV-CH(FV)-FV	CH	1

CONSTRAINTS

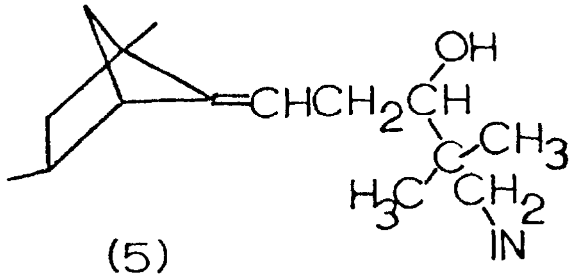
1) ALL FREE VALENCES BONDED TO NON-HYDROGEN ATOMS

2) GOODLIST
IN-CH2-BI 1 TO ANY
(EVENTUALLY IN-CH_2-$CH_{0 \to 0}$)
ME-(BI CH) 1 TO ANY
(EVENTUALLY CH_3-CH, EXACTLY 1)

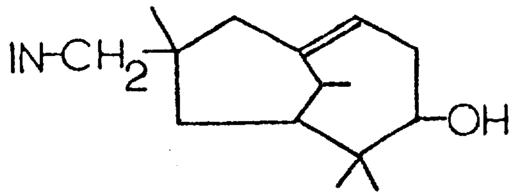
3) GOODRINGS 2 EXACTLY 5

4) BADRINGS 3

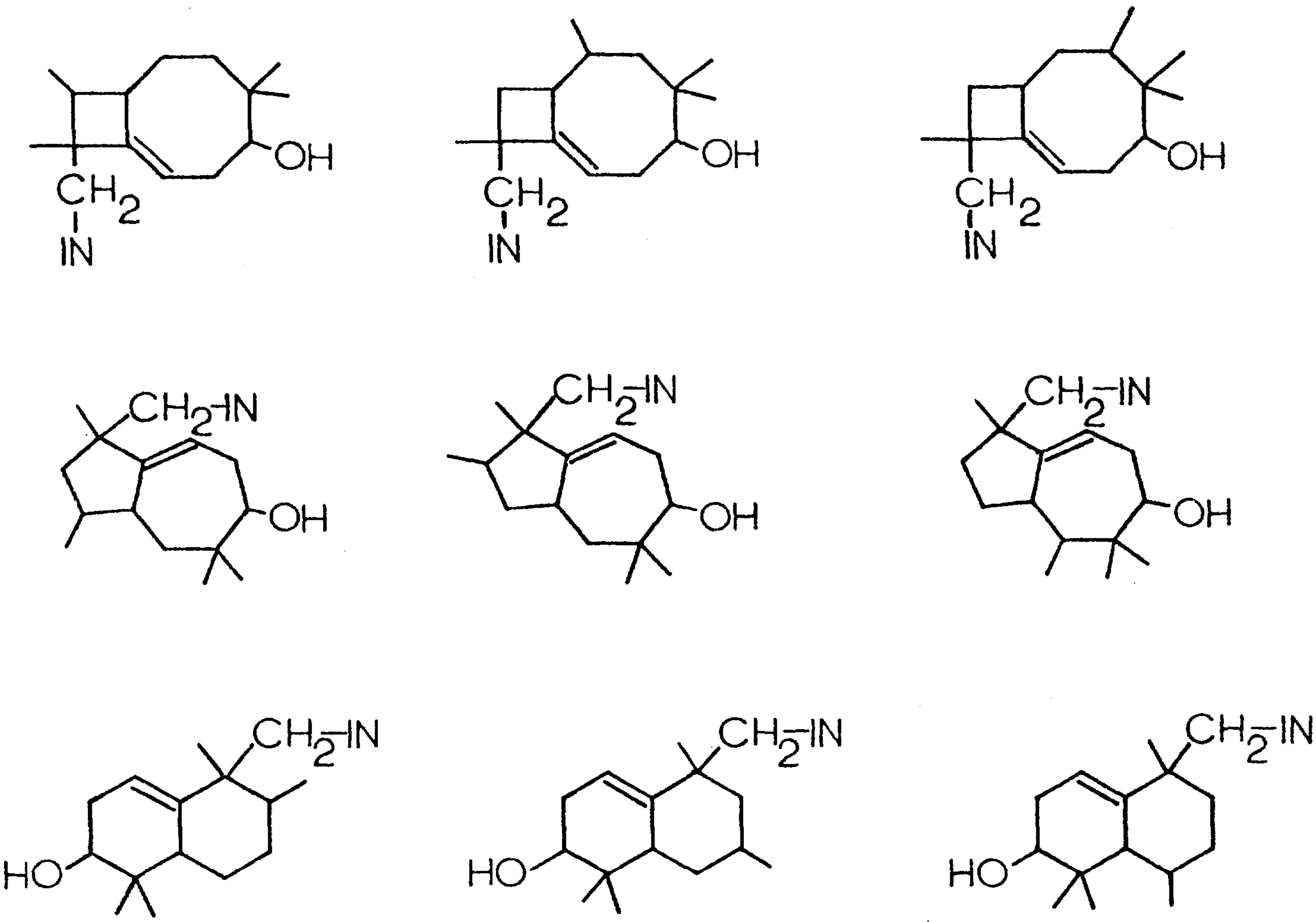
Figure 4. Superatoms and constraints supplied to CONGEN in investigations of plausible structural alternatives to the proposed structure of Polyalthenol.



(5)



(6)



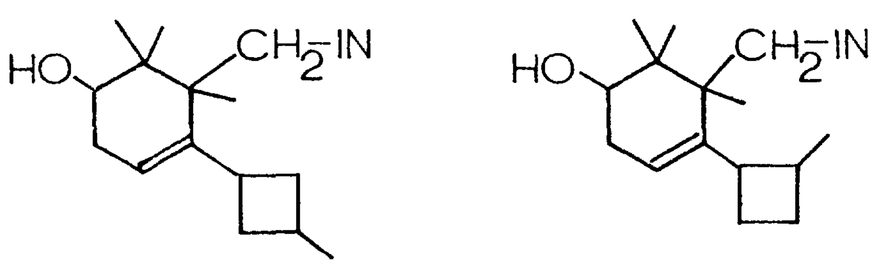


Figure 5.

Structural candidates for polyalthenol based on data given in Figure 4.

REACTION CHEMISTRY DEVELOPMENTS

1. SEPARATION FROM CONGEN - COMMUNICATION VIA FILES OF STRUCTURES.

2. ADDING CONSTRAINTS - SITE - AND TRANSFORM - SPECIFIC.

3. CONTROL STRUCTURE - RAMIFICATION

   A. ESTABLISH RELATIONSHIPS AMONG PRODUCTS AND REACTANTS
   B. DEAL PROPERLY WITH RANGES OF NUMBERS OF PRODUCTS

4. INTERACTION - DEVELOP MANIPULATION COMMANDS WHICH PARALLEL LABORATORY OPERATIONS, E.G., SEPARATE INTO FLASKS, TEST CONTENTS OF VARIOUS FLASKS, INCOMPLETE SEPARATIONS, ETC.

5. REPRESENTATION OF REACTIONS

6. PROSPECTIVE DETECTION OF DUPLICATE PRODUCTS BASED ON SYMMETRY PROPERTIES OF: A) STARTING MATERIAL; AND B) TRANSFORMATION.

Figure 6. Current and future direction for improvement and extension of REACT, a program for exploration of applications of reaction chemistry to structure elucidation problems.